Computational identification of noncoding RNAs in E. coli by comparative genomics

Elena Rivas, Robert J. Klein, Thomas A. Jones & Sean R. Eddy

Address: Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, Saint Louis, Missouri 63110, USA.

Correspondence: Sean R. Eddy

E-mail: eddy@genetics.wustl.edu

Some genes produce noncoding transcripts that function directly as structural, regulatory, or even catalytic RNAs [1,2]. Unlike protein coding genes, which can be detected as open reading frames with distinctive statistical biases, noncoding RNA (ncRNA) gene sequences have no obvious inherent statistical biases [3]. Thus, genome sequence analyses reveal novel protein-coding genes, but any novel ncRNA genes remain invisible. Here we describe a computational comparative genomic screen for ncRNA genes. The key idea is to distinguish conserved RNA secondary structures from a background of other conserved sequences using probabilistic models of expected mutational patterns in pairwise sequence alignments. We report the first whole-genome screen for ncRNA genes done with this method, in which we applied it to the "intergenic" spacers of *Escherichia coli* using comparative sequence data from four related bacteria. Starting from >23,000 conserved interspecies pairwise alignments, the screen predicted 275 candidate structural RNA loci. A sample of 49 candidate loci was assayed experimentally. At least eleven expressed small, apparently noncoding RNA transcripts of unknown function. Our computational approach may be used to discover structural ncRNA

genes in any genome for which appropriate comparative genome sequence data are available.

Results and discussion

Comparative genome analysis reveals important sequences that cannot be detected in a single genome analysis [4]. A standard comparative approach uses pairwise alignment methods to find significantly conserved sequences. However, a bewildering number of sequences are conserved for many different reasons, including coding exons, a variety of transcriptional and post-transcriptional *cis*-regulatory signals, and sequences of as yet unknown function. If the genomic region of interest can be localized *a priori*, one can deal with a manageable subset of conserved sequences, as is the case for predicting transcriptional *cis*-regulatory sites upstream of known genes by "phylogenetic

-8]. To screen an entire genome for a specific, unusual type of conserved sequence, such as ncRNA genes, a more specific approach is desirable.

The key idea of our approach for distinguishing a small number of structural RNAs from a large background of other conserved sequences is to exploit a distinctive *pattern* of mutation, as opposed to simple conservation, as sketched in Figure 1. A conserved structural RNA tends to show a pattern of compensatory mutations consistent with some base-paired secondary structure. A conserved coding region tends to show a pattern of synonymous codon substitutions [9,10]. Other types of conserved region may be approximated by a "null hypothesis" that mutations occur position-independently with no special pattern. Thus, we should be able to use position-specific mutational models to separate conserved regions into three types: probable structural RNAs, probable coding regions, and probable "other" sequences.

Badger and Olsen [9] described a two-model protein genefinding approach that distinguishes coding alignments from conserved noncoding alignments. We extended this

idea by including a third model of RNA structure evolution, by allowing the input alignments to be gapped, and by using full Bayesian probabilistic models. Our three probabilistic models (RNA, COD, and IND) are stochastic "pair grammars": a pair stochastic context free grammar (SCFG) as the RNA model, and pair hidden Markov models for both COD and IND [11]. The secondary structure model in the pair-SCFG is a full probabilistic analogue of the Zuker MFOLD algorithm [11], including terms for base stacking, loop lengths, etc. that have been trained on a database of tRNA and rRNA secondary structures [3]. The evolutionary distance component of all three models is derived directly (for COD) or indirectly (for RNA and IND) from a single choice of amino acid substitution matrix (BLOSUM62) [13,14]. It is essential that all three models are at the same evolutionary distance. Otherwise, models might distinguish alignments solely based on their level of conservation rather than the pattern of mutation.

Given an input pairwise sequence alignment (for instance, from a BLASTN comparison of two related genomes), we score the alignment with each of the three models. The scoring algorithms are all dynamic programming algorithms; the rate-limiting one is the $O(L^2)$ time, $O(L^3)$ memory pair-SCFG alignment algorithm used to score the RNA model, which must be done as an SCFG Inside algorithm [11], summing over all possible RNA secondary structures. The scores are log likelihoods that are then used to calculate a final log-odds score for the RNA model compared to the other two models. Non-RNA alignments get negative scores; increasingly positive scores indicate increasingly strong comparative evidence that the alignment contains a conserved structural RNA.

A satisfactory mathematical description of the approach is beyond the scope of this paper, and will be published elsewhere (E.R. and S.R.E., manuscript submitted). We have implemented the complete approach in a program, QRNA [15]. The goal of this brief communication is to describe the first whole-genome screen we have done with QRNA.

We used QRNA to screen the complete genome of *Escherichia coli* K12 MG1655 (version M52) [16]. We chose *E. coli* because there is a complete genome sequence, it is a simple model genetic system in which we could readily test our predictions experimentally, and there is extensive comparative sequence coverage from at least fifteen related species [17]. We chose four related enterobacterial genomes for comparison: *Salmonella typhi* [18], *S. paratyphi* A [19], *S. enteriditis* [20], and *Klebisella pneumoniae* [19].

We started with all annotated intergenic sequences with a length 50 nt in *E. coli* according to U. Wisconsin's annotation of 115 ncRNA genes and 4290 coding ORFs [21]. This gave 2367 intergenic sequences, totaling about 500 kb. The average sequence length was 211 nt; the longest was 1729 nt. Four known ncRNA genes (csrB, oxyS, micF, and rprA) [2,22] were unannotated and were left in the dataset as positive controls.

Each "intergenic" region was used as a query to search the four comparative genome databases using BLASTN [23,24] (2.0MP-WashU, 12 Feb 01 version, default parameters and scoring matrix). All alignments with an E-value <0.01, a length of 50 nt, and an overall identity of 65% were collected, giving a database of 23,674 pairwise BLASTN alignments (12,037 from *S. typhi*, 5,239 from *S. paratyphi*, 4,260 from *S. enteriditis*, and 2,138 from *K. pneumoniae*).

Each pairwise alignment was analyzed with QRNA, in "local Viterbi" mode (where it finds a locally optimal RNA structure, allowing the BLASTN alignment to extend into other conserved sequence flanking the RNA), scanning the pairwise alignment in overlapping 200 nt windows moving 50 nt at a time. All windows classified as "RNA" by the program with a log odds score of 5 bits were kept and overlapping windows were merged. (Annotated ncRNAs in *E. coli* score between 5.6 and 41.1 bits. Non-RNA alignments rarely score above 0.) This resulted in 556 candidate RNA loci. All four positive controls were classified as RNA (5.6 for micF, 5.7 for oxyS, 10.6 for csrB, and

13.6 for rprA). The COD class also detected 160 candidates for conserved small ORFs that were not examined further.

The complete screen took about 20 hours on a single SGI Origin200 R10K processor. A variety of tests on QRNA's performance suggest that under the conditions above, it has a sensitivity of about 80% on known structural RNAs (E.R. and S.R.E., manuscript submitted). As a test of specificity, we shuffled each of the 23,674 input alignments by aligned columns (preserving % identity, while scrambling the sequences and any position-specific mutational pattern in the alignment) and applied the same procedure, which produced 73 false positive "RNA" loci with scores over 5 bits. Therefore about 85% of our 556 candidate loci should be "true positives", in the sense that they are not just the result of expected statistical noise.

Because QRNA screens for conserved RNA secondary structure, we expected it to detect various nongenic sequences with conserved RNA structure, including rho-independent terminators, rRNA spacers, transcriptional attenuators in ribosomal protein and amino acid biosynthetic operons [25], other *cis*-regulatory RNA structures [26], and even certain repetitive elements [27,28]. We removed 281 loci that plausibly fell into one of these nongenic classes, leaving a total of 275 candidate loci (see Supplementary Table 1 for a list). It must also be noted that not all ncRNA genes conserve an intramolecular secondary structure - for example, QRNA does not detect C/D box small nucleolar RNAs [29] in yeast or *Pyrococcus* alignments.

We expected these 275 loci to be a mix of ncRNA genes, *cis*-regulatory RNA structures, and false positives. A criterion that tends to distinguish an ncRNA gene from either *cis*-regulatory structures or false positive signals is the expression of a distinct transcript independent of adjacent coding genes. 49 candidate loci were assayed for expression by Northern blotting (see Supplementary Table 2 for a list). For 11 loci, we observed discrete small RNA transcripts < 400 nt (Figure 2). Six others showed larger products

that were interpreted as coding mRNA transcripts. Two showed multiple discrete bands that we could not interpret. The remaining 30 loci were not expressed under these growth conditions. We cannot interpret a negative result because several known ncRNAs are expressed only under specific conditions (for example, OxyS RNA is expressed under oxidative stress conditions but not in normal lab growth [30]). The 11 loci that express small RNAs are listed in Table 1.

Expression is suggestive, but not entirely sufficient to define a candidate as a new ncRNA gene. For example, the *cis*-acting transcriptional attenuator of the *his* operon [31] is detected by QRNA, is flanked by a strong consensus promoter and an obvious rho-independent terminator, has no significant coding potential (the *hisL* leader peptide is only 22 aa long), and Northern analysis detects the *his* leader RNA as a distinct 170 nt transcript (data not shown). Candidates t44, tpk1, and tpk2 are directly upstream of coding genes in the same orientation, and may be attenuators. We also cannot exclude the possibility that expressed candidates are protein-coding genes with small ORFs. For example, candidate k4 is classified as RNA in a *Klebsiella* alignment but as coding in *Salmonella* alignments; the sequence appears to contain both a 72 aa conserved ORF with a reasonable translational initiation consensus and a conserved structural RNA motif, overlapping each other. The semantic concept of a “gene” is slippery to begin with, especially when the gene is noncoding. Therefore, although we conclude from our Northern assays that a significant number of our 275 candidate loci do indeed correspond to independent ncRNA genes, each individual candidate will require detailed study. We have deliberately not assigned any new *E. coli* gene names to our loci at this point, pending more complete experimental characterization that is ongoing in our lab.

While this paper was in preparation, two groups reported exciting results of different screens for small ncRNAs in the intergenic regions of the *E. coli* genome sequence [32,33]. Wassarman *et al.* used sequence conservation coupled with microarray

expression analysis and found 17 new ncRNAs [32]. Argaman *et al.* used sequence conservation coupled with promoter and rho-independent terminator prediction and found 14 new ncRNAs [33]. Both groups report extensive experimental characterization of the new loci. The overlap of these experimentally confirmed ncRNA genes with our results gives us additional confidence in QRNA's sensitivity. Ten of the fourteen RNAs reported by Argaman *et al.*, are in our list of 275 candidate loci; of the four that we do not detect (sraD, sraH, sraI, sraL), three have scores only slightly below our 5 bit cutoff, and only sraI was completely missed. Fourteen of the seventeen RNAs reported by Wassarman *et al.* were detected by QRNA; of the three that we missed (ryeA, ryhA, and ryjA), two were just below our cutoff, and one (ryeA) was detected in the initial list of 556 QRNA candidates but we mistakenly discarded it, thinking it was just a rho-independent terminator. On the other hand, only 4/11 of our confirmed candidates were detected and confirmed by one of the other screens (Table 1), which suggests that QRNA's sensitivity is higher than either the Argaman *et al.* or the Wassarman *et al.* screens, and that neither of these screens saturated the *E. coli* genome for novel ncRNAs.

These data, though experimentally preliminary, nonetheless validate QRNA as a powerful and general means for identifying candidate structural ncRNA loci. Because we use no organism-specific information (such as promoter or terminator consensus sequences), QRNA will be applicable in any organism for which appropriate comparative genomic data are available. We have already anecdotally observed that signal/noise is sufficient to screen the human genome using low-pass mouse shotgun sequence coverage: for example, a QRNA screen of the 196 kb draft sequence of a human BAC (Genbank accession AL357874) spanning the cartilage hair hypoplasia (CHH) locus [34], using unassembled 1.7X mouse shotgun coverage (Mouse Sequencing Consortium, unpublished), predicts two ncRNA loci (data not shown), one of which corresponds to the 265 nt RNase MRP ncRNA locus that has recently been implicated as the gene

responsible for CHH [35]. Given the recent surge in comparative genome sequencing, we will be able to screen for structural ncRNAs in all the major systems, including human (by comparison to mouse), *Caenorhabditis elegans* (via *C. briggsae*), *Drosophila melanogaster (*via *D. pseudoobscura*), *Saccharomyces cerevisiae (*via multiple other yeast genomes), and *Arabidopsis thaliana* (via *Brassica*).

Acknowledgements

This work was supported by the Howard Hughes Medical Institute, the NIH National Human Genome Research Institute, a Sloan Foundation postdoctoral fellowship to E.R., and an HHMI graduate fellowship to R.J.K.

References

1. Eddy SR: **Noncoding RNA genes.** *Curr Opin Genet Dev* 1999, **9**:695-699.

2. Wassarman KM, Zhang A, Storz G: **Small RNAs in *Escherichia coli*.** *Trends Microbiol* 2000, **7**:37-45.

3. Rivas E, Eddy SR: **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics* 2000, **16**:583-605.

4. Miller W: **Comparison of genomic DNA sequences: solved and unsolved problems.** *Bioinformatics* 2001, **17**:391-397.

5. Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome.** *Genome Res* 1997, **7**:959-966.

6. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26**:225-228.

7. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, *et al*.: **Active conservation of noncoding sequences revealed by three-way species comparisons.** *Genome Res* 2000, **10**:1304-1306.

8. Pennacchio LA, Rubin, EM: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2**:100-109.

9. Badger JH, Olsen, GJ: **CRITICA: Coding region identification tool invoking comparative analysis.** *Mol Biol Evol* 1999, **16**:512-524.

10. Crollius HR, Jaillon O, Dasilva C, Bouneau L, Fischer C, Fizames C, *et al.*: **Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence.** *Nat Genet* 2000, **25**:235-238.

11. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge: Cambridge University Press; 1998.

12. Zuker M: **Calculating nucleic acid secondary structure.** *Curr Opin Struct Biol* 2000, **10**:303-310.

13. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89**:10915-10919.

14. Altschul SF: **Amino acid substitution matrices from an information theoretic perspective.** *J Mol Biol* 1991, **219**:555-565.

15. ftp://ftp.genetics.wustl.edu/pub/eddy/software/qrna.tar.Z

16. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1474.

17. http://www.tigr.org/

18. Unpublished data from Beowulf Genomics and the Pathogen Sequencing Group at the Sanger Centre; ftp://ftp.sanger.ac.uk/pub/pathogens/st/

19. Unpublished data from the Washington University Genome Sequencing Center; http://genome.wustl.edu/gsc/Projects/bacteria.shtml

20. Unpublished data from the Department of Microbiology at the University of Illinois; http://www.salmonella.org/

21. http://www.genome.wisc.edu/

22. Majdalani N, Chen S, Murrow J, St. John K, Gottesman S: **Regulation of RpoS by a novel small RNA: the characterization of RprA.** *Mol Microbiol* 2001, **39**:1382-1394.

23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

24. http://blast.wustl.edu/

25. Yanofsky C: **Transcription attenuation: once viewed as a novel regulatory strategy.** *J Bacteriol* 2000, **182:**1-8.

26. Storz G: **An RNA thermometer.** *Genes Dev* 1999, **13:**633-636.

27. Cai XY, Maxon ME, Redfield B, Glass R, Brot N, Weissbach H: **Methionine synthesis in *Escherichia coli*: effect of the MetR protein on metE and metH expression.** *Proc Natl Acad Sci USA* 1989, **86:**4407-4411.

28. Bachellier S, Clément JM, Hofnung M, Gilson E: **Bacterial interspersed mosaic elements (BIMEs) are a major source of sequence polymorphism in *Escherichia coli* intergenic regions including specific association with a new insertion sequence.** *Genetics* 1999, **145:**551-562.

29. Lowe TM, Eddy SR: **A computational screen for methylation guide snoRNAs in yeast.** *Science* 1999, **283:**1168-1171.

30. Altuvia S, Weinstein-Fischer D, Zhang A, Postow L, Storz G: **A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator.** *Cell* 1997, **90:**43-53.

31. Chan CL, Landick R: **The *Salmonella typhimurium* his operon leader region contains an RNA hairpin-dependent transcription pause site. Mechanistic implications of the effect on pausing of altered RNA hairpins.** *J Biol Chem* 1989, **264:**20796-20804.

32. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S: **Identification of novel small RNAs using comparative genomics and microarrays.** *Genes Dev* 2001, **15**:1637-1651.

33. Argaman L, Hershberg R, Vogel J, Bejerano G, Gerhart E, Wagner H, *et al.*: **Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli.*** *Curr Biol* 2001, **11**:941-950.

34. Ridanpää M, van Eenennaam H, Pelin K, Chadwick R, Johnson C, Yuan B, *et al.*: **Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia.** *Cell* 2001, **104**:195-203.

35. Rudd KE: **Novel intergenic repeats of *Escherichia coli* K-12.** *Res Microbiol* 1999, **150**:653-664.

Table 1. Candidate loci expressing small RNAs

Candidate	Strand	Transcript size, nt	QRNA score, bits	Candidate start	Candidate size, nt	Adjacent ORFs		Note	Gene
tpke11	+	370	19.5	14,080	88	dnaK>	dnaJ>	Overlaps ORF(s)	
tp2	-	60; 120	25.0	122,857	160	pdhR>	aceE>		
t44	+	135	8.0	189,678	195	map<	rpsB>	Attenuator?	
tpe7	-	65	21.3	1,762,685	190	sufA<	ydiH<		rydB
tpke70	-	40	12.3	2,494,649	435	lpxP>	ypdZ<		
tp1	+	300	29.1	2,651,802	373	sseA>	sseB<	PAIR3 [34]	ryfA
tke1	+	150; 180	19.5	2,689,183	212	yfhK<	purL<		
tp8	-	110; 140	18.3	3,192,705	254	yqiK>	rfaE<	QUAD1d [34]	
tpk1	+	120; 180	37.9	3,235,948	274	ygjR>	ygjT>	Attenuator?	sraF
k4	-	200	19.3	3,436,082	197	mscL>	zntR<	72 aa ORF (yhdL)	
tpk2	+	250	25.2	4,048,659	268	yihA<	yihI>	Attenuator?	sraK,ryiB

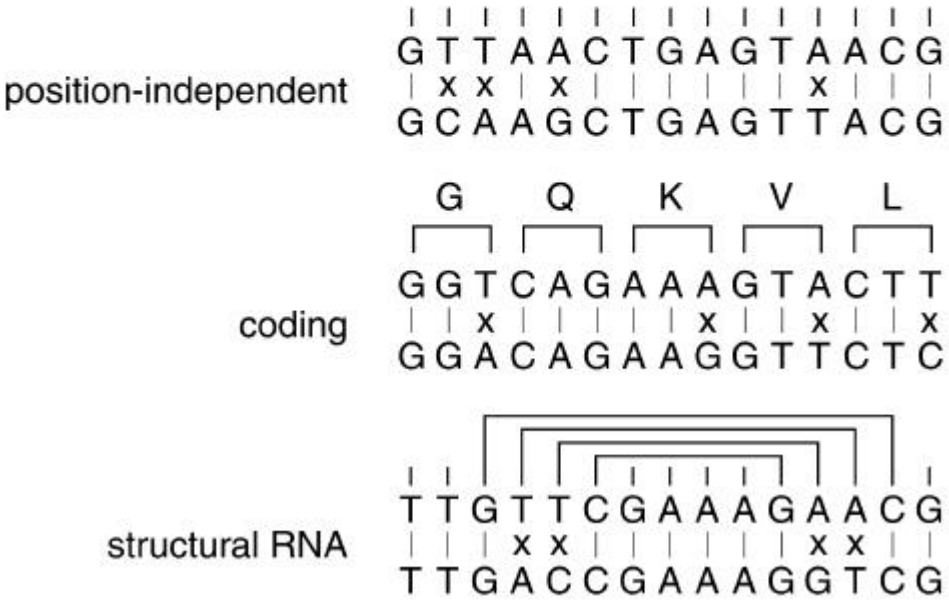
Candidate names are arbitrary codes, not final gene names. Candidate positions are for the computational prediction, not the observed transcript, and are relative to the M52 version of the genome. Transcriptional direction of adjacent ORFs is indicated by > or <. The "gene" column indicates gene names assigned to candidates that were also experimentally confirmed by either Argaman et al. [33] or Wassarman et al.[32].

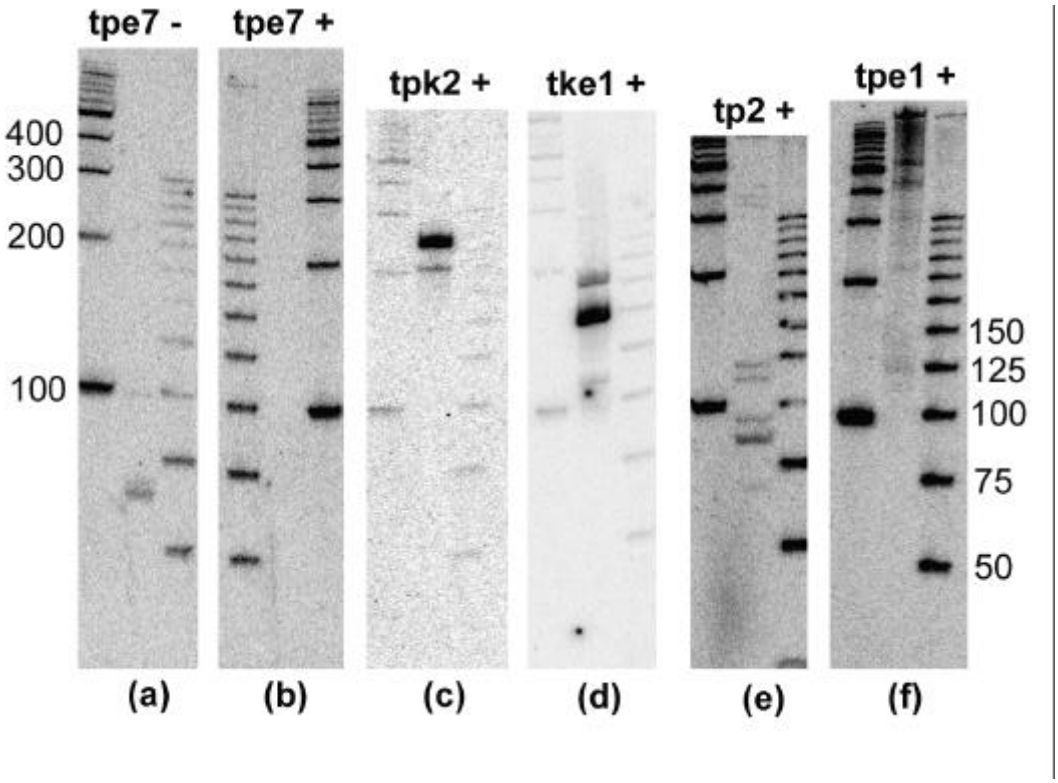
Figure legends.

Figure 1. Three pairwise alignments of identical composition with identical number and type of base substitutions can be classified by distinctive patterns of mutation caused by different selective constraints: the position-independent null hypothesis (top), a coding region (middle), or a structural RNA (bottom). Marks above each alignment indicate how it is scored to calculate each model's likelihood : one position at a time for IND, one codon at a time for COD (integrated over all six possible frames), and as a combination of base-paired positions and single positions for RNA (integrated over all possible secondary structures).

Figure 2. Total *E. coli* RNA was isolated from log-phase cultures growing in rich (LB) medium at 37°C, run on a 6% polyacrylamide gel, electroblotted, and probed with a 5'-^{32}P labeled oligonucleotide specific to each strand of the predicted locus. Panels a and b show a typical positive result: tpe7 hybridizes to a 65 nt RNA transcript only with a - strand probe (a), and not with a + strand probe (b). Panels c, d, and e: confirmed transcripts for candidates tpk2, tke1, and tp2. Panel f: a typical result for a candidate (tpe1, downstream of *rpsA*) that more likely is part of an mRNA.

eddy_fig1


position-independent
G T T A A C T G A G T A A C G
x x x x
G C A A G C T G A G T T A C G
G Q K V L
coding
G G T C A G A A A G T A C T T
x x x x x
G G A C A G A A G G T T C T C
structural RNA
T T G T T C G A A A G A A C G
x x x x
T T G A C C G A A A G G T C G


tpe7 -
tpe7 +
tpk2 +
tke1 +
tp2 +
tpe1 +
400
300
200
100
150
125
100
75
50
(a)
(b)
(c)
(d)
(e)
(f)